ANNUAL REVIEWS Further
Click here to view this article's online features:
- Download figures as PPT slides
- Navigate linked references
- Download citations
- Explore related articles
- Search keywords

Mediation Analysis: A Practitioner's Guide

Tyler J. VanderWeele

T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts 02115;
email: tvanderw@hsph.harvard.edu

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

Annu. Rev. Public Health 2016. 37:17–32

First published online as a Review in Advance on November 30, 2015

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

This article's doi:
10.1146/annurev-publhealth-032315-021402

Copyright © 2016 by Annual Reviews.
All rights reserved

Keywords

direct effects, indirect effects, mechanism, pathway analysis

Abstract

This article provides an overview of recent developments in mediation analysis, that is, analyses used to assess the relative magnitude of different pathways and mechanisms by which an exposure may affect an outcome. Traditional approaches to mediation in the biomedical and social sciences are described. Attention is given to the confounding assumptions required for a causal interpretation of direct and indirect effect estimates. Methods from the causal inference literature to conduct mediation in the presence of exposure-mediator interactions, binary outcomes, binary mediators, and case-control study designs are presented. Sensitivity analysis techniques for unmeasured confounding and measurement error are introduced. Discussion is given to extensions to time-to-event outcomes and multiple mediators. Further flexible modeling strategies arising from the precise counterfactual definitions of direct and indirect effects are also described. The focus throughout is on methodology that is easily implementable in practice across a broad range of potential applications.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

INTRODUCTION

Methodology for mediation to assess the importance of various pathways and mechanisms has expanded dramatically over the past decade. The topic of mediation has traditionally been more in the provenance of social scientists and psychologists, and training and education on methodological approaches for mediation have been less common in epidemiology and public health. Many of the recent methodologic advances have, however, come out of the causal inference, biostatistics, and epidemiologic research communities. This review takes the reader through what some of these advances have been, with an eye toward those developments that may be useful in the practice of epidemiology and public health research. A full book-length overview of these topics is now available (47), and the present review in some sense serves as a guide to that fuller treatment of the subject. A similar review is also available on the topic of interaction (48). The present article provides a succinct overview of methodology for mediation; describes when and in which contexts traditional approaches are valid and in which settings other analytic approaches need to be sought out; and points to relevant papers and relevant sections of the book-length overview for further reading.

TRADITIONAL APPROACHES TO MEDIATION ANALYSIS

Investigators have utilized two traditional approaches to mediation analysis, sometimes referred to respectively as the difference method and the product method (or product-of-coefficients method). We consider each in turn.

The Difference Method

The difference method has been employed more frequently in epidemiology and the biomedical sciences. It consists of fitting two regression models. Let A denote an exposure of interest, M a potential mediator, Y an outcome, and C a set of baseline covariates. The first regression model for the difference method is simply a regression of the outcome Y on the exposure A and covariates C:

$$E[Y|a, c] = \phi_0 + \phi_1 a + \phi_4' c.$$

The coefficient, ϕ_1, is interpreted as the total effect of the exposure A on the outcome Y. The second regression is similar but includes the mediator as a variable in the regression as well:

$$E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c.$$

If the exposure coefficient of the first regression, ϕ_1, without the mediator, goes down considerably when comparing it with the exposure coefficient in the second regression, θ_1, when adding the mediator, this is thought to be indicative of mediation because it seems as though the mediator explains some of the effect of the exposure on the outcome. The difference between these two coefficients is sometimes interpreted as a mediated or indirect effect (IE):

$$IE = \phi_1 - \theta_1.$$

The exposure coefficient itself, θ_1, in the model that includes the mediator is then generally taken as a measure of the direct effect (DE) because the effect on the outcome appears to remain even when control has been made for the mediator:

$$DE = \theta_1.$$

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

We return in the next section to the assumptions under which the interpretation of these estimates as direct and indirect effects is valid.

The Product Method

A slightly different method, sometimes called the product method or the product-of-coefficients method, is used with more frequency in the social sciences. The approach was made popular in part by a paper by Baron & Kenny (5), though it had been proposed earlier (3, 14, 20, 33). With the product method, two regressions are again employed. We once again regress the outcome on the exposure, the mediator, and the covariates:

$$E[Y \mid a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4' c.$$

We then regress the mediator itself on the exposure and the covariates:

$$E[M \mid a, c] = \beta_0 + \beta_1 a + \beta_2' c.$$

The direct effect is once again taken as θ_1, the exposure coefficient in the outcome regression model that includes the mediator. The indirect effect, however, is taken as the product of β_1 and θ_2, i.e., the exposure coefficient in the mediator model times the mediator coefficient in the outcome model. The product $\beta_1\theta_2$, taken as a measure of the indirect effect, thus has a seemingly intuitive interpretation as the effect of the exposure on the mediator times the effect of the mediator on the outcome.

With these two methods, the product method and the difference method, the question naturally arises about how they compare. Fortunately, for a continuous outcome and mediator with linear regression models fit by ordinary least squares, the two approaches coincide. Numerically we will always have (25) for our mediated effect that $\beta_1\theta_2 = \phi_1 - \theta_1$. However, this is not the case with logistic regression models. With logistic regression with a binary outcome, the product and difference methods do not give numerically identical results. We return below to this question of which, if either, of these methods for logistic regression is valid.

CONFOUNDING ASSUMPTIONS FOR MEDIATION ANALYSIS

Fairly strong assumptions are needed for the estimates of direct and indirect effects to be interpreted causally. First, as in ordinary observational studies, control must be made for exposure-outcome confounding (Assumption A1). Second, because with direct and indirect effects we are also drawing conclusions about the effects of the mediator on the outcome, control must be made for mediator-outcome confounding (Assumption A2). Third, because mediation analysis is essentially about the exposure changing the mediator (and that change in the mediator affecting the outcome), control must also be made for exposure-mediator confounding (Assumption A3). Finally, for standard estimates, as above, to be interpreted as direct and indirect effects, there should be no mediator-outcome confounder that is itself affected by the exposure (Assumption A4). We now consider each of these confounding assumptions in more detail.

Graphically, one might picture the first three assumptions as in **Figure 1**. These three assumptions—control for exposure-outcome, mediator-outcome, and exposure-mediator confounding—essentially amount to controlling for the variables C_1, C_2, and C_3 in **Figure 1**, corresponding with exposure-outcome confounders, mediator-outcome confounders, and exposure-mediator confounders, respectively. In practice, some of the covariates may affect all the exposure, mediator, and outcome, and the covariates may also affect each other. None of this is problematic and the covariate groups C_1, C_2, and C_3 need not be distinguished from one another. What is

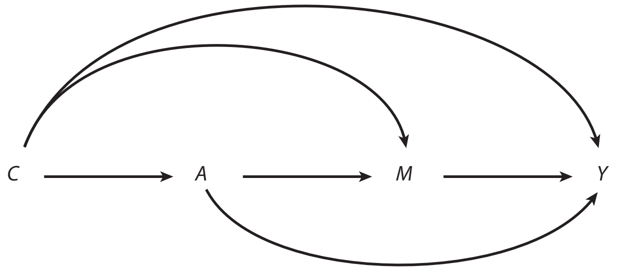

Figure 1

Relations between exposure *A*, mediator *M*, and outcome *Y*, and confounders.

important is that the covariates included in the regression models above suffice to control for exposure-outcome, mediator-outcome, and exposure-mediator confounding (Assumptions A1–A3). These are, of course, strong assumptions; in a later section, we consider sensitivity analysis, which allows us to assess how robust our direct and indirect effect estimates are to violations in these assumptions and how substantial a violation in the assumptions would have to be in order to considerably alter our inferences about direct and indirect effects.

We have thus far considered the first three of the assumptions described above. The fourth assumption is that none of the mediator-outcome confounders are themselves affected by the exposure. Diagrammatically, in **Figure 1**, this assumption essentially corresponds to there being no arrow from the exposure *A* to the mediator-outcome confounder C_2. If the exposure did affect a mediator-outcome confounder C_2, then this would be problematic for the estimation of direct and indirect effects because the variable C_2 would then itself also be a mediator for the effect of the exposure *A* on the outcome *Y* and one that itself also confounded the effect of the mediator of interest, *M*, on the outcome *Y*. The fourth assumption would thus be violated in a setting such as that depicted in **Figure 2** because in this figure L affects both the mediator *M* and outcome *Y* and is itself affected by exposure *A*. Addressing scenarios such as this (when another variable is both a mediator and a confounder for our mediator of interest, *M*) is more complicated. We consider this setting in a later section in which we discuss concepts and methods for handling multiple mediators. For the next several sections, however, we assume that this fourth assumption of no mediator-outcome confounders affected by the exposure also holds. Another way to think about the fourth assumption is that there should be relatively little time between the exposure

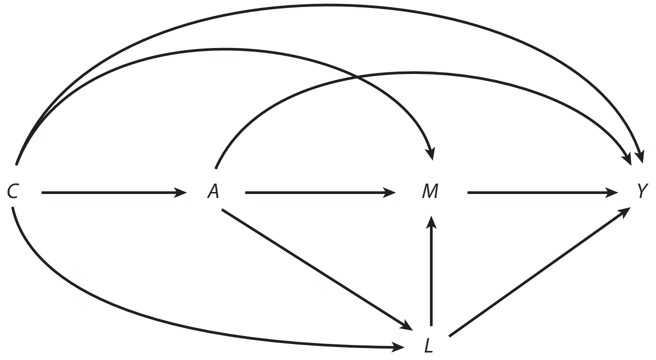

Figure 2

A mediator-outcome confounder *L* that is itself affected by the exposure *A*.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

and the mediator. If a substantial gap exists, then the fourth assumption requires that nothing on the pathway from the exposure to the mediator itself also independently affects the outcome. This assumption likely becomes increasingly less plausible the more time that elapses between the exposure and the mediator.

The confounding assumptions for mediation analysis are extremely important. Violations in these assumptions can give rise to very misleading results. Assumptions A1 (control for exposure-outcome confounding) and A3 (control for exposure-mediator confounding) correspond to assumptions typically made in observational studies for total effects. What distinguishes the assumptions required in the mediation context is that control must also be made for mediator-outcome confounding (Assumption A2). This assumption is not necessary for the analysis of total effects, but it is needed for the analysis of direct and indirect effects. Moreover, to estimate direct and indirect effects, this assumption is needed even if the exposure has been randomized. The assumption of control for mediator-outcome confounding is not needed for the analysis of total effects in a randomized trial, but it is needed for the analysis of direct and indirect effects. It is needed even in the randomized trial context because, in a trial, although the exposure has been randomized, the mediator typically has not been. Once we start reasoning about direct and indirect effects, we are considering the effects of not only the exposure but also the mediator as well. Once again, failure to control for mediator-outcome confounding in a randomized trial can substantially bias estimates of direct and indirect effects.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

As an example of such bias, Strong et al. (34) consider the effects of a randomized cognitive behavioral therapy intervention on depressive symptoms at three months follow-up and found there to be a beneficial effect: Those who had received the therapy had lower depressive symptoms in follow-up. However, those in the therapy arm had, at three months follow-up, higher rates of antidepressant use. This observation led to questions regarding whether the cognitive behavioral therapy intervention had a beneficial effect on depressive symptoms only because it resulted in higher antidepressant use or whether the intervention affected depressive symptoms through other pathways by changing thought and behavioral patterns. If the intervention were beneficial only because of higher antidepressant use, then the cognitive-behavioral aspects of the intervention could be abandoned and a more cost-effective intervention, focusing only on antidepressant adherence, could be employed. To address this question, if we apply the traditional approaches to mediation analysis above to the Strong et al. (34) data and regress depressive symptoms on antidepressant use and therapy, then the coefficient for antidepressant use in this regression is positive (47). With a naïve analysis, it may appear that antidepressants increase depression. The effect mediated by antidepressant use thus looks to be harmful, and the direct effect is larger than the total effect. In short, we get nonsense from the traditional approach if we ignore mediator-outcome confounding. What is almost certainly occurring here is that those using antidepressants are likely also those in more difficult contexts, e.g., those who are having relationship troubles or who have lost loved ones. The confounding between antidepressant use and depressive symptoms (i.e., mediator-outcome confounding) is so severe that we even get the direction of the regression coefficient for antidepressant use wrong. See VanderWeele (47, section 3.4) for further discussion of this example and reanalysis of the data.

Once again, this highlights the fact that if we are interested in mediation analysis, then we must control for mediator-outcome confounders. This necessity applies to the traditional approaches to mediation analysis described above, and it also applies to more recent methods described below. With either the traditional approaches or with the more recent approaches, for direct and indirect effects estimates to have a causal interpretation, we need Assumptions A1–A4. If researchers were to more regularly design studies so as to collect data on potential mediator-outcome confounders and adjust for these in mediation analysis, inferences about direct and indirect effects would be

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

more likely to be valid. In a subsequent section, we consider sensitivity analysis for such mediator-outcome confounding.

ALLOWING FOR EXPOSURE-MEDIATOR INTERACTION

In addition to clarifying the confounding assumptions required to estimate direct and indirect effects, the causal inference literature on mediation has also clarified how mediation analysis can be conducted and how a total effect can be decomposed into direct and indirect effects, even when the exposure and the mediator interact in their effects on the outcome. Suppose, for example, that our model for the outcome included an exposure-mediator interaction,

$$E[Y \mid a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c,$$

and that we once again fit a linear regression model for the mediator as was the case with the product method,

$$E[M \mid a, c] = \beta_0 + \beta_1 a + \beta_2' c.$$

If the models are correctly specified and the confounding Assumptions A1–A4 hold, then direct and indirect effect estimates for a change in the exposure from level a to a^* (e.g., for a binary exposure $a = 1, a^* = 0$) are given by (52)

$$DE = \{\theta_1 + \theta_3(\beta_0 + \beta_1 a^* + \beta_2' c)\}(a - a^*)$$
$$IE = (\beta_1\theta_2 + \beta_1\theta_3 a)(a - a^*).$$

Standard errors for these expressions are also available (52), and software to compute these effects and their standard errors and confidence intervals automatically are available in SAS (Statistical Analysis System), Stata, and SPSS (Statistical Package for the Social Sciences) (39). The total effect is the sum of the direct and indirect effects, and sometimes a proportion-mediated measure is used, obtained by dividing the indirect effect by the total effect [see VanderWeele (47, section 2.13) for further discussion of the measure and its properties].

Note that when there is no exposure-mediator interaction (i.e., when $\theta_3 = 0$), these expressions simply reduce to θ_1 for the direct effect and $\beta_1\theta_2$ for the indirect effect, i.e., the same estimates as those of the product method described above (which also with linear regression coincides with the difference method). The other terms in the expressions for the direct and indirect effects account for the presence of exposure-mediator interaction.

BINARY OUTCOMES AND LOGISTIC REGRESSION

A similar approach to that described in the previous section can be applied with a binary outcome and logistic regression. Suppose we have a binary outcome and a normally distributed continuous mediator, and we fit a logistic regression model for the outcome, possibly allowing for exposure-mediator interaction,

$$logit\{P(Y = 1 \mid a, m, c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c,$$

and that we once again fit a linear regression model for the mediator,

$$E[M \mid a, c] = \beta_0 + \beta_1 a + \beta_2' c.$$

Provided the outcome is relatively rare (a point to which we return below) and provided that confounding Assumptions A1–A4 hold and the models are correctly specified, then direct and indirect effect estimates on an odds ratio scale are given approximately by (53)

$$\begin{aligned}\log\{OR^{DE}\} &\cong \{\theta_1 + \theta_3(\beta_0 + \beta_1 a + \beta_2' c + \theta_2\sigma^2)\}(a - a^*) + 0.5\theta_3^2\sigma^2(a^2 - a^{*2})\\ \log\{OR^{IE}\} &\cong (\theta_2\beta_1 + \theta_3\beta_1 a)(a - a^*),\end{aligned}$$

where σ^2 is the variance of the error term in the regression for the mediator. Once again, standard errors for these expressions are also available (53), and software to compute these effects and their standard errors and confidence intervals automatically are available in SAS, Stata, and SPSS (39). For a rare outcome, a proportion-mediated measure can be obtained from the direct and indirect effect odds ratios by the formula $\mathrm{OR}^{\mathrm{DE}}(\mathrm{OR}^{\mathrm{IE}}-1)/(\mathrm{OR}^{\mathrm{DE}}\mathrm{OR}^{\mathrm{IE}}-1)$.

If there is no exposure-mediator interaction (i.e., if $\theta_3 = 0$), these expressions simply reduce to $\exp(\theta_1)$ for the direct effect and $\exp(\beta_1\theta_2)$ for the indirect effect. Thus, we once again have product method-type expressions. We had noted above that the product method and the difference method do not coincide for logistic regression. However, if the outcome is rare (and there is no interaction in the model) then the product method and difference method do at least approximately coincide (53).

However, if the outcome is not rare (10% is often used as a cutoff), then the product method and the difference method can and do diverge, and, in fact, neither of these approaches nor the expressions given above are valid for the direct and indirect effects. We return to this issue shortly. One way around the problem if the outcome is common is to replace the logistic regression model above with a log-binomial model. If we use a log-binomial model for the outcome, then the expressions above are valid for the direct and indirect effects on a risk ratio scale and, in the absence of interaction, the product and difference methods will coincide. Software for estimating direct and indirect effects with log-binomial models is also available (39).

The problem with binary outcomes that are not rare pertains to the fact that logistic regression uses the odds ratio, which is a measure that is "noncollapsible" (11), and thus marginal and conditional odds ratios are not directly comparable. With a common outcome, the odds ratios with the mediator in the model versus the odds ratios without the mediator are thus not directly comparable, which leads to problems with the difference method. The problem arises because as we add covariates to the logistic regression model (even if these are not confounders), the coefficients tend to increase in magnitude (cf. 32). When the outcome is common, the odds ratio does not approximate the risk ratio, and the extent of this lack of approximation can vary with the other covariates in the models.

If we add the mediator to the logistic regression outcome model, then the coefficient of the exposure in the logistic regression may go down somewhat because of mediation but go up somewhat because of the additional variable in the model. It might, then, appear as though the coefficient of the exposure does not change at all even though there is, in fact, mediation. We would then draw the wrong conclusion from the difference method. In fact, because of this noncollapsibility of odds ratios, it can be shown that, with logistic regression, the difference method is conservative for mediation. If one uses the difference method and if the confounding assumptions hold, the difference method will generally underestimate the indirect effect when used with logistic regression (19). Thus if the difference method with logistic regression indicates the presence of a mediated effect, then there is, in fact, evidence for a mediated effect. Yet, if the difference method does not indicate a nonzero estimate of the indirect effect, one cannot assume that there is no mediation; there may still be mediation, but the difference method does not allow one to draw conclusions in this case because the difference method is conservative. When the outcome is rare, odds ratios approximate risk ratios and these problems vanish. When the outcome is common, we can circumvent these issues by fitting a log-linear model rather than a logistic model. Further discussion is given in VanderWeele (47, chapter 2).

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

One final point is worth noting when using binary outcomes. Our discussion so far is relevant for estimating direct and indirect effects with cohort data. Often in epidemiologic studies, data are available from a case-control study. With a case-control study design in which sampling is done on the basis of the outcome Y, the estimates from the logistic regression model for that outcome can be used in the analysis. However, the regression model for the mediator needs to be modified to account for the sampling design. To do so, one can use a weighting technique (53), or, if the outcome is rare, a much simpler approach is to fit the mediator model only among the controls. With a rare outcome, the distribution of the mediator among the controls will be a very close approximation to the distribution of the mediator in the underlying population; therefore, the direct and indirect effect estimates with the mediator model fit only among the controls will give a very close approximation to the direct and indirect effects. Software for estimating direct and indirect effects for a case-control design is also available (39).

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

BINARY MEDIATORS

A similar regression-based approach to estimating direct and indirect effects is also applicable to binary mediators. Suppose that the mediator is binary and the outcome is continuous and that the following models are correctly specified, which allows for exposure-mediator interaction:

$$\begin{aligned} E[Y = 1|a, m, c] &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ logit\{P(M = 1|a, c)\} &= \beta_0 + \beta_1 a + \beta_2' c\,. \end{aligned}$$

If the models are correctly specified and the confounding Assumptions A1–A4 hold, then direct and indirect effect estimates are given by

$$\begin{aligned} DE &= \theta_1(a - a^*) + \theta_3(a - a^*)\frac{\exp(\beta_0 + \beta_1 a^* + \beta_2' c)}{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)} \\ IE &= (\theta_2 + \theta_3 a)\left\{\frac{\exp(\beta_0 + \beta_1 a + \beta_2' c)}{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)} - \frac{\exp(\beta_0 + \beta_1 a^* + \beta_2' c)}{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)}\right\}. \end{aligned}$$

Once again these are simply a combination of the regression coefficients of the two regression models. Likewise, if both the mediator and the outcome are binary and we fit two logistic regression models

$$\begin{aligned} logit\{P(Y = 1|a, m, c)\} &= \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c \\ logit\{P(M = 1|a, c)\} &= \beta_0 + \beta_1 a + \beta_2' c\,, \end{aligned}$$

then, if the models are correctly specified and the confounding Assumptions A1–A4 hold, then the direct and indirect effects are given by

$$\begin{aligned} OR^{DE} &\cong \frac{\exp(\theta_1 a)\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}{\exp(\theta_1 a^*)\{1 + \exp(\theta_2 + \theta_3 a^* + \beta_0 + \beta_1 a^* + \beta_2' c)\}} \\ OR^{IE} &\cong \frac{\{1 + \exp(\beta_0 + \beta_1 a^* + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a + \beta_2' c)\}}{\{1 + \exp(\beta_0 + \beta_1 a + \beta_2' c)\}\{1 + \exp(\theta_2 + \theta_3 a + \beta_0 + \beta_1 a^* + \beta_2' c)\}}. \end{aligned}$$

Software to estimate direct and indirect effects with binary mediators is also available (39).

SENSITIVITY ANALYSIS FOR UNMEASURED CONFOUNDING

We have emphasized above that the assumptions needed to draw conclusions about direct and indirect effects, Assumptions A1–A4, are quite strong and will often be violated in applications. It is therefore important to assess how robust one's conclusions are about direct and indirect effects to violations in the assumptions being made. Sensitivity analysis techniques help one assess, for

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

example, how strong an unmeasured confounder would have to be related to both the mediator and to the outcome to substantially change conclusions being drawn about the direct and indirect effects. Several sensitivity analysis techniques for mediation have been proposed in the literature [12, 15, 35, 42; see chapter 3 of VanderWeele (47) for an overview of many of these methods].

Here we focus on one very recent sensitivity analysis technique (8) that has broad applicability and makes relatively few assumptions and is thus of potential use in a wide range of potential applications. The technique uses two sensitivity analysis parameters. It assumes that exposure-outcome confounding (Assumption A1) and exposure-mediator confounding (Assumption A3) are controlled but that there might be unmeasured mediator-outcome confounders (i.e., Assumption A2 is violated), although it assumes that none of these is affected by the exposure (i.e., Assumption A4 is satisfied). Thus, we consider a causal structure such as that in **Figure 3**. Consider a binary outcome and let *U* be an unmeasured mediator-outcome confounder and let

$$\gamma = \max m \frac{\max_u P(Y = 1|A = 1, m, c, u)}{\min_u P(Y = 1|A = 1, m, c, u)}$$

denote the maximum risk ratio relating *U* and the outcome *Y* among the exposed subjects across strata of the mediator, conditional on covariates. This parameter is the maximum ratio by which *U* can increase the likelihood of the outcome *Y* via pathways other than through *M*. Second, let

$$\lambda = \max_{u,m} \frac{P(u|A = 1, m, c)}{P(u|A = 0, m, c)}$$

denote the maximum risk ratio relating *U* and the exposure across different conditional levels of *M*. This second parameter is somewhat more difficult to interpret. If both *A* and *U* affect *M*, then conditional on *M*, there will be an association between *A* and *U* even if neither affects the other. We are specifying the maximum of these associations. Some intuition for this parameter can be gained in noting that over many scenarios (10) the association between *A* and *U* within strata of *M* will generally be smaller than the magnitude of the ratio association between *A* and *M* and also smaller than that between *U* and *M*, and in the opposite direction. Thus, these can be helpful in specifying possible values of the second sensitivity analysis parameter λ.

It can be shown (8) that if **Figure 3** represents the correct causal structure, then the maximum ratio that such an unmeasured confounder can decrease the direct effect or increase the indirect

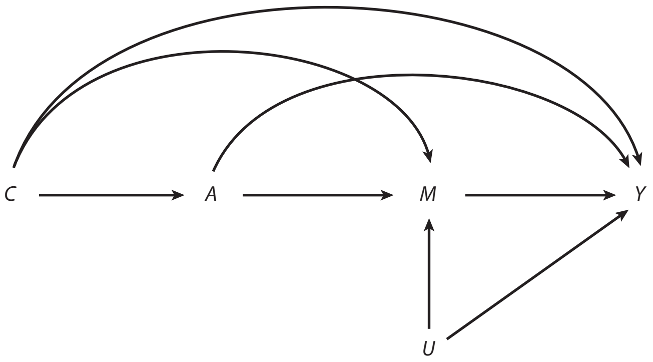

Figure 3

An unmeasured confounder *U* of the mediator-outcome relationship.

effect is given by

$$B = \frac{\gamma\lambda}{\gamma + \lambda - 1}.$$

Thus, to get a corrected estimate (understood as the most that such an unmeasured confounder can alter the direct and indirect effects) we can take our direct and indirect effects estimates and their confidence intervals from the observed data and divide the direct effect estimate and both limits of its confidence interval by the bias factor B and then multiply the indirect effect estimate and both limits of its confidence interval by the bias factor B to obtain such corrected estimates. The actual mechanics of this technique are fairly simple in practice. The technique holds under the assumption that **Figure 3** is the correct causal structure, i.e., that we have only unmeasured mediator-outcome confounding and that our other confounding Assumptions (A1, A2, and A4) hold.

We do not know, of course, what the sensitivity analysis parameters, γ and λ, are, but we can vary them to see how large they must be before estimates change in meaningful ways or, for example, are reduced to the null. In practice, reporting an entire table of corrected direct and indirect effect estimates across a whole range of sensitivity analysis parameters can be helpful. This can be done by putting increasingly large values of one sensitivity analysis parameter λ on the rows and the other parameter γ on the columns and reporting the corrected estimates and confidence intervals for each of these settings. Doing so gives the reader considerable information on how sensitive estimates are to violations in the assumptions. At the very least, however, it is good practice to report how much unmeasured confounding would be required to reduce the direct effect estimate to the null and also how much confounding would be required to reduce to confidence interval to include the null. This can all be done in a relatively straightforward manner.

The technique we have described here is relevant for binary outcomes. However, similar techniques can be used for continuous, count, and time-to-event outcomes and can also be applied on a difference rather than a ratio scale. The reader is referred to Ding & VanderWeele (8) for further discussion of the technique.

MEASUREMENT ERROR AND MISCLASSIFICATION

Recent work has addressed the impact of measurement error and misclassification on direct and indirect effect estimates. Several correction methods are now available (18, 23, 40, 38, 51) that employ regression calibration techniques, SIMEX (Simulation Extrapolation) methods, methods of moments estimators, weighting approaches, and the EM (Expectation-Maximization) algorithm. Other work has also considered differential measurement error of the mediator (23). Software is available to implement some of these various techniques, but resources are still somewhat limited. The reader is referred to the relevant papers (18, 17, 40, 38, 51) for further information or to VanderWeele (47, section 3.5). Here we focus on some additional intuitive results concerning the direction of the bias subject to nondifferential measurement error of the mediator, or exposure, or outcome.

We begin with the potential measurement error or misclassification of the mediator. If the mediator is binary or if both the mediator and outcome are continuous, and there is no exposure-mediator interaction, then one can show (28, 51) that the indirect effect will be biased toward the null and the direct effect will be biased away from the null. The intuition here is that measurement error or misclassification of the mediator will weaken the association between the mediator and the

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

outcome. As the indirect effect can often be thought of as the product of the effect of the exposure on the mediator and that of the mediator on the outcome, this indirect effect will be biased downward by the measurement error weakening the association between the mediator and the outcome. In addition, because the indirect effect is biased toward the null, the direct effect will be biased away from the null. This intuition always holds either if the mediator is binary or if both the mediator and outcome are continuous and if there is no exposure-mediator interaction. It will often hold in other scenarios (e.g., if the mediator has three or more levels or is continuous with exposure-mediator interaction), but it will not always hold in these other scenarios. Correction techniques (40, 38) can still be used in these other scenarios if it is unclear whether the intuition applies.

Other work has considered the biases of direct and indirect effect estimators in the presence of nondifferential measurement error of the exposure or the outcome (18, 17). For nondifferential measurement error of the outcome, both direct and indirect effects are unbiased for continuous outcomes, and both are biased toward the null for dichotomous outcomes (18). Drawing intuitive conclusions about the direction of the bias of direct and indirect effects is thus relatively straightforward in the context of measurement error of the outcome.

For nondifferential measurement error of the exposure, in the absence of exposure-mediator interaction, the natural direct effect is biased toward the null, but the indirect effect can be biased in either direction (17). The intuition for the indirect effect is that measurement error of the exposure will tend to weaken the exposure-mediator association but will strengthen the mediator-outcome association. Which of these two consequences is more substantial will determine whether the indirect effect is biased toward or away from the null. Jiang & VanderWeele (18, 17) also developed correction methods for direct and indirect effects estimators in the presence of nondifferential measurement error of the exposure and the outcome.

TIME-TO-EVENT OUTCOMES

A similar approach to mediation analysis can be used with time-to-event outcomes as well. Fuller discussion is given elsewhere (47), but one can once again specify, for example, either a proportional hazard model or an accelerated failure time model for a time-to-event outcome and either a linear or logistic regression for a continuous or binary mediator, respectively. The coefficients can again be combined to obtain estimates of direct and indirect effects. As was the case with logistic regression, so also with proportional hazards models, the product and the difference methods require a rare outcome assumption (e.g., less than 10% by the end of follow-up) to be applicable (43). The use of accelerated failure time models does not require this assumption. For a proportional hazards model with a common outcome, a weighting approach can be used (22). Methods for mediation with time-to-event outcomes have also been developed using additive hazard models (21). Sensitivity analysis techniques are also available (45). Further discussion of mediation with time-to-event outcomes can be found in chapter 4 of VanderWeele (47).

MULTIPLE MEDIATORS

Our discussion thus far has concerned only a single mediator. Methods are also available for multiple mediators. Sometimes an informal approach is used for multiple mediators by assessing mediation one mediator at a time and then summing the proportion mediated across mediators. If the mediators affect one another, then this approach fails. Even if the mediators do not affect one another, this approach will still fail if there are interactions between the effects of the various mediators on the outcome. A regression-based approach, similar to that described above, for assessing the extent to which the effect of an exposure is mediated by an entire set of mediators can be used to address these settings in which the mediators might affect one another (cf. 54); these methods can be used even if the ordering of the mediators is unknown. A weighting-based approach can be used for even greater flexibility (54). VanderWeele (47, chapter 5) provides additional information and the precise assumptions and methodology required.

A more challenging setting is to assess the effect mediated through one intermediate when there are other mediators that precede and affect the mediator of interest, such as in **Figure 2** above. In this context, direct and indirect effects are generally not identified, even if one has data on all the variables (4, 47), unless one makes further strong modeling assumptions about linearity of the models and the absence of certain interactions (7, 16, 36, 57), as is done in a linear structural equation model. Some progress can be made in this context using sensitivity analysis (16, 36, 57). As discussed above, in this context one can still assess the effects mediated jointly by the mediator of interest as well as those preceding it. Certain path-specific effects can also be estimated (47, 55). While structural equation models (SEMs) allow one to assess many effects, they also make assumptions about linearity and normality for all variables in the model and require that the relations between all variables are unconfounded (44). Essentially, for an SEM, the confounding assumptions A1–A4 that are described above are needed not just for a single exposure, mediator, or outcome, but for every set of variables on the SEM. These are very strong assumptions and will often not hold. SEMs thus deliver more effects than do the methods described in this article, but they require much stronger assumptions. They can be useful for hypothesis generation but need to be interpreted cautiously.

PRECISE COUNTERFACTUAL INTERPRETATION

The recent progress that has been made in methods for mediation has come about through approaching the question of mediation from a counterfactual-based perspective on causal inference. We briefly describe here the counterfactual definitions of direct and indirect effects that have allowed this. Let Y_a denote a subject's outcome if exposure A were set, possibly contrary to fact, to a. Let M_a denote a subject's counterfactual value of the intermediate M if exposure A were set to the value a. Finally, let Y_{am} denote a subject's counterfactual value for Y if A were set to a and M were set to m. Robins & Greenland (30) and Pearl (29) gave the following definitions for controlled direct effects and natural direct and indirect effects based on interventions on the mediator M. The controlled direct effect of exposure A on outcome Y comparing $A = a$ with $A = a^*$ and setting M to m is defined by $Y_{am} - Y_{a^*m}$ and measures the effect of A on Y not mediated through M—that is, the effect of A on Y after intervening to fix the mediator to some value m. In contrast with controlled direct effects, natural direct effects fix the intermediate variable for each individual to the level which it naturally would have been under—for example, the absence of exposure. The natural direct effect of exposure A on outcome Y comparing $A = a$ with $A = a^*$ intervening to set M to what it would have been if exposure had been $A = a^*$ is formally defined by $Y_{aM_{a^*}} - Y_{a^*M_{a^*}}$. Corresponding to a natural direct effect is a natural indirect effect formally defined by $Y_{aM_a} - Y_{aM_{a^*}}$. The natural indirect effect assumes that exposure is set to some level $A = a$ and then compares what would have happened if the mediator were set to what it would have been if exposure had been a versus what would have happened if the mediator were set to what it would have been if exposure had been a^*.

Under the confounding Assumptions A1–A4, these effects are identified on average for a population by the methods and expressions shown above. The controlled direct effects require only Assumptions A1 and A2. Controlled direct effects can also be estimated even when there are

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

mediator-outcome confounders affected by the exposure as in **Figure 2**, although special methods such as marginal structural models and structural nested models are then needed (41, 56) because the regression methods above will no longer suffice. For further discussion of Assumptions A1–A4, their interpretation, and some of the controversies concerning their interpretation, the reader is directed to Robins & Richardson (31) and VanderWeele (47, sections 2.3 and 7.3).

In the absence of exposure-mediator interaction in the models above, the controlled direct effects are equal to the natural direct effects. Controlled direct effects cannot in general be used for effect decomposition or to assess the relevance of a particular pathway (there is generally no "controlled indirect effect"). However, controlled direct effects are often of greater policy relevance because they consider the effect of the exposure that would remain under an intervention on the mediator to fix it to a specific value. Sometimes a proportion eliminated measure is reported that is defined as the difference between the total and controlled direct effect, divided then by the total effect; the measure will differ from the proportion mediated in the presence of exposure-mediator interaction. See VanderWeele (47, sections 2.13 and 2.14) for further discussion.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

MORE FLEXIBLE MODELS

In this review, we have considered various parametric models to undertake mediation analysis. However, the causal inference approach to mediation is very flexible and can be pursued under any model. The difficulty is that each time the model is changed, new expressions for the effects have to be derived. The SAS, Stata, and SPSS macros described above (39) consider numerous different scenarios. However, if greater flexibility is desired, a simulation-based approach has been developed by Imai et al. (15), which allows investigators to specify much more flexible models for the outcome and the mediator and then to estimate the direct and indirect effects by simulation. The approach makes the same confounding Assumptions A1–A4 but allows for more flexible modeling. Software is available in both R and Stata (6, 37). For further discussion, the reader is directed to Imai et al. (15), Tingley et al. (37), or VanderWeele (47, sections 2.17 and 2.18).

CONCLUSION

Mediation analysis has expanded rapidly over the past decade. Numerous other methods have been developed (1, 2, 9, 13, 24, 26, 35, 58), which we could not address in this article. Some of these are discussed in the book length treatment of mediation (47). Methods have begun to be developed for handling questions of mediation for time-varying exposures and mediators (50), but more work remains to be done in this area. These ideas have also found application in health disparities research (e.g., 27, 49). Concepts and methods are now also available to assess mediation and interaction simultaneously. A total effect can, in fact, be decomposed into not just two but four distinct components: the effect due only to mediation, that due only to interaction, that due to both mediation and interaction, and that due to neither (46). The new methodology considers how much of the direct effect described above is or is not also due to interaction and how much of the indirect effect described above is or is not also due to interaction. The approach provides maximum insight into the phenomena of mediation and interaction simultaneously. SAS code to implement this type of analysis is also available (46). See VanderWeele (46; 47, chapter 14) for further discussion. Methodology continues to advance, and further developments in the years ahead are likely.

The mediation methods discussed in this review can be useful for a number of purposes. Some potential uses of these ideas and methods include trying to understand etiology, providing evidence to confirm and refute theory, assessing the impact of intervening on a mediator when it is not possible to alter an exposure, and trying to understand why an intervention succeeded or failed. The application of these techniques makes some strong assumptions and should thus always be accompanied by sensitivity analysis; in at least some instances, these approaches can give considerable insight into pathways. Further discussion of motivations for and uses of mediation analysis is given in VanderWeele (47, section 1.3).

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The research was supported by NIH grant ES017876.

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

LITERATURE CITED

1. Albert JM. 2012. Distribution-free mediation analysis for nonlinear models with confounding. *Epidemiology* 23:879–88
2. Albert JM, Nelson S. 2011. Generalized causal mediation analysis. *Biometrics* 67:1028–38
3. Alwin DF, Hauser RM. 1975. The decomposition of effects in path analysis. *Am. Sociol. Rev.* 40:37–47
4. Avin C, Shpitser I, Pearl J. 2005. Identifiability of path-specific effects. *Proc. Int. Jt. Conf. Artif. Intel., Edinburgh, Aug.*, pp. 357–63
5. Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51:1173–82
6. Daniel RM, De Stavola BL, Cousens SN. 2011. gformula: estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata J.* 11:479–517
7. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. 2015. Causal mediation analysis with multiple mediators. *Biometrics* 71:1–14
8. Ding P, VanderWeele TJ. 2015. Sharp sensitivity bounds for mediation under unmeasured mediator-outcome confounding. *Tech. Rep.* In press
9. Goetgeluk S, Vansteelandt S, Goetghebeur E. 2008. Estimation of controlled direct effects. *J. R. Stat. Soc. Ser. B Stat Methodol.* 70:1049–66
10. Greenland S. 2003. Quantifying biases in causal models: classical confounding versus collider-stratification bias. *Epidemiology* 14:300–6
11. Greenland S, Robins JM, Pearl J. 1999. Confounding and collapsibility in causal inference. *Stat. Sci.* 14:29–46
12. Hafeman DM. 2011. Confounding of indirect effects: a sensitivity analysis exploring the range of bias due to a cause common to both the mediator and the outcome. *Am. J. Epidemiol.* 174:710–17
13. Hong G, Nomi T. 2012. Weighting methods for assessing policy effects mediated by peer change. *J. Res. Educ. Eff.* (Spec. Issue: *Stat. Approaches Stud. Mediat. Eff. Educ. Res.*) 5:261–89
14. Hyman HH. 1955. *Survey Design and Analysis: Principles, Cases and Procedures*. Glencoe, IL: Free Press
15. Imai K, Keele L, Tingley D. 2010. A general approach to causal mediation analysis. *Psychol. Methods* 15:309–34
16. Imai K, Yamamoto T. 2012. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Pol. Anal.* 21:141–71
17. Jiang Z, VanderWeele TJ. 2015. Causal mediation analysis in the presence of a misclassified binary exposure. *Tech. Rep.*
18. Jiang Z, VanderWeele TJ. 2015. Causal mediation analysis in the presence of a mismeasured outcome. *Epidemiology* 26:e8–9

19. Jiang Z, VanderWeele TJ. 2015. When is the difference method conservative for mediation? *Am. J. Epidemiol.* 182(2):105–8

20. Judd CM, Kenny DA. 1981. Process analysis: estimating mediation in treatment evaluations. *Eval. Rev.* 5:602–19

21. Lange T, Hansen JV. 2011. Direct and indirect effects in a survival context. *Epidemiology* 22:575–81

22. Lange T, Vansteelandt S, Bekaert M. 2012. A simple unified approach for estimating natural direct and indirect effects. *Am. J. Epidemiol.* 176:190–95

23. le Cessie S, Debeij J, Rosendaal FR, Cannegieter SC, Vandenbroucke J. 2012. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology* 23:551–60

24. MacKinnon DP. 2008. *Introduction to Statistical Mediation Analysis*. New York: Erlbaum

25. MacKinnon DP, Warsi G, Dwyer JH. 1995. A simulation study of mediated effect measures. *Multivar. Behav. Res.* 30:41–62

26. Martinussen T, Vansteelandt S, Gerster M, von Bornemann Hjelmborg J. 2011. Estimation of direct effects for survival data by using the Aalen additive hazards model. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73:773–88

27. Nandi A, Glymour MM, Kawachi I, VanderWeele TJ. 2012. Using marginal structural models to estimate the direct effect of adverse childhood social conditions on onset of heart disease, diabetes and stroke. *Epidemiology* 23:223–32

28. Ogburn EL, VanderWeele TJ. 2012. Analytic results on the bias due to nondifferential misclassification of a binary mediator. *Am. J. Epidemiol.* 176:555–61

29. Pearl J. 2001. Direct and indirect effects. *Proc. Conf. Uncertain. Artif. Intel., 17th, Seattle*, pp. 411–20. San Francisco: Kaufmann

30. Robins JM, Greenland S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3:143–55

31. Robins JM, Richardson TS. 2010. Alternative graphical causal models and the identification of direct effects. In *Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, ed. P Shrout, pp. 103–58. Oxford, UK: Oxford Univ. Press

32. Robinson L, Jewell NP. 1991. Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* 59:227–40

33. Sobel ME. 1982. Asymptotic confidence intervals for indirect effects in structural equations models. In *Sociological Methodology*, ed. S Leinhart, pp. 290–312. San Francisco: Jossey-Bass

34. Strong V, Waters R, Hibberd C, Murray G, Wall L, et al. 2008. Management of depression for people with cancer (SMaRT oncology 1): a randomised trial. *Lancet* 372:40–48

35. Tchetgen Tchetgen EJ, Shpitser I. 2012. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann. Stat.* 40:1816–45

36. Tchetgen Tchetgen EJ, VanderWeele TJ. 2014. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology* 25:282–91

37. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. 2014. Mediation: R package for causal mediation analysis. *J. Stat. Softw.* 59:1–38

38. Valeri L, Lin X, VanderWeele TJ. 2014. Mediation analysis when a continuous mediator is measured with error and the outcome follows a generalized linear model. *Stat. Med.* 33:4875–90

39. Valeri L, VanderWeele TJ. 2013. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol. Methods* 18:137–50

40. Valeri L, VanderWeele TJ. 2014. The estimation of direct and indirect causal effects in the presence of a misclassified binary mediator. *Biostatistics* 15:498–512

41. VanderWeele TJ. 2009. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 20:18–26

42. VanderWeele TJ. 2010. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* 21:540–51

43. VanderWeele TJ. 2011. Causal mediation analysis with survival data. *Epidemiology* 22:575–81

44. VanderWeele TJ. 2012. Structural equation modeling in epidemiologic analysis. *Am. J. Epidemiol.* 176:608–12

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.

45. VanderWeele TJ. 2013. Unmeasured confounding and hazard scales: sensitivity analysis for total, direct and indirect effects. *Eur. J. Epidemiol.* 28:113–17

46. VanderWeele TJ. 2014. A unification of mediation and interaction: a four-way decomposition. *Epidemiology* 25:749–61

47. VanderWeele TJ. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford Univ. Press

48. VanderWeele TJ, Knol MJ. 2014. A tutorial on interaction. *Epidemiol. Methods* 3:33–72

49. VanderWeele TJ, Robinson W. 2014. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 25:473–84

50. VanderWeele TJ, Tchetgen Tchetgen EJ. 2014. *Mediation analysis with time-varying exposures and mediators*. Work. Pap. 168, Harvard Univ. Biostat. Work. Pap. Ser. **http://biostats.bepress.com/cgi/viewcontent.cgi?article=1176&context=harvardbiostat**

51. VanderWeele TJ, Valeri L, Ogburn EL. 2012. The role of misclassification and measurement error in mediation analyses. *Epidemiology* 23:561–64

52. VanderWeele TJ, Vansteelandt S. 2009. Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* 2:457–68

53. VanderWeele TJ, Vansteelandt S. 2010. Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.* 172:1339–48

54. VanderWeele TJ, Vansteelandt S. 2013. Mediation analysis with multiple mediators. *Epidemiol. Methods* 2:95–115

55. VanderWeele TJ, Vansteelandt S, Robins JM. 2014. Methods for effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology* 25:300–6

56. Vansteelandt S. 2009. Estimating direct effects in cohort and case-control studies. *Epidemiology* 20:851–60

57. Vansteelandt S, VanderWeele TJ. 2012. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics* 68:1019–27

58. Zheng W, van der Laan MJ. 2012. Targeted maximum likelihood estimation of natural direct effects. *Int. J. Biostat.* 8:1–40

Annu. Rev. Public Health 2016.37:17-32. Downloaded from www.annualreviews.org
Access provided by McGill University on 08/16/16. For personal use only.